CHAPTER 1: THINKING TIME-SERIALLY

This chapter defines time series analysis and distinguishes time series data from other forms of data. It provides an introduction to some notation and terminology that will set the framework for the discussion of time series fundamentals in Chapter 2 and to two of the examples that will be used throughout the text. It will end with a discussion of some of the opportunities and challenges of time series analysis, which will be expanded on in subsequent chapters.

1.1 Time Series Analysis and Time Series Data

Time series analysis in the social sciences is the application of statistical models to time series data to examine the movement of social science variables over time (e.g., public opinion, government policy, judicial decisions, educational outcomes, socioeconomic measures), allowing analysts to estimate relationships within (over time) and between variables in order to test causal hypotheses, make forecasts about the future, and assess the impact of policy changes.

To clarify exactly what time series data are and are not, it is useful to compare such data to other types of data. For many, the most familiar type of data is cross-sectional data. Typically, cross-sectional data are from a random sample of cases. For example, a variable Y is a collection of observations on randomly selected cases:

$$Y = \{y_1, y_2, y_3, \ldots, y_N\}, N = \text{number of cases.} \quad (1.1.1)$$

Each observation y_i is from a different case, all from the same point in time. If the cases are selected by simple random sampling, each value of y_i is roughly independent of the others. Cases can be a random selection of individuals, countries, firms, and so on.

As an example of cross-sectional data, consider the following cross-sectional data on individual preferences for total government spending in 1976 in Britain (Table 1.1).

Cross-sectional data have one observation for each case. Time series data have a separate observation for each time point, and each observation is for the same case—for example, GDP (gross domestic product) of a country. The time between observations can be years, months, days, hours, and so on. However, as we shall see, the measurements are assumed to be

Table 1.1 Individual Preferences for Total Government Spending in 1976 Britain

Individual	*Spending Preference*
1	2
2	2
3	1
4	5
5	4
6	5
7	4
8	3
9	5
10	5
11	3
12	4
13	2
14	5
⋮	⋮

NOTE: Individual preferences for total government spending: scored *strongly in favor* (1), *in favor* (2), *neither in favor nor against* (3), *against* (4), or *strongly against* (5) government spending cuts.

(roughly) evenly spaced. A time series variable Y_t is a nonrandom sequence of observations for an individual case ordered over time:

$$Y_t = \{y_1, y_2, y_3, \ldots, y_T\},\ T = \text{number of time points.} \qquad (1.1.2)$$

Again as an example, consider the following time series data on *net* preferences for total government spending in Great Britain each year from 1975 onward (Table 1.2).

Another type of data, which is neither cross-sectional nor purely time series, is panel data. Continuing our previous example, consider the following data on net preferences for total government spending in a selection of countries at three time points: 1986, 1996, and 2006. Panel data can be presented in either stacked or nonstacked format (Table 1.3).

Table 1.2 Net Preferences for Total Government Spending in Great Britain

Year	*Spending Preference*
1975	−6.1
1976	−7.9
1977	−9.4
1978	−10.1
1979	−10.7
1980	−11.5
1981	−12.4
1982	−12.6
1983	−12.2
1984	−13.5
1985	−14.8
1986	−14.4
1987	−14.4
1988	−14.3
⋮	⋮

NOTE: Net preferences for total government spending: the average of survey responses, scored strongly in favor (−100), in favor (−50), neither in favor nor against (0), against (+50), or strongly against (+100) government spending cuts. The measure ranges in theory from −100, meaning that all respondents strongly favor spending cuts, to +100, meaning that all respondents oppose spending cuts.

In panel data, we have more than one case. The same set of cases is observed at multiple time points. Typically in the social sciences, we observe more cases than we do time points.

A final type of data is pooled cross-sections. This is not quite panel data as the cases measured at each time point are not the same. For example, our data may be the responses of individuals from Britain to a monthly survey, with a different random sample of individuals each month. This type of data is often collapsed into a time series for the

Table 1.3 Net Preferences for Total Government Spending

Nonstacked Country	*1986*	*1996*	*2006*
Australia	–45.3	–37.3	–12.8
Germany	–51.8	–67.5	–54.9
Great Britain	–14.4	–12.8	–4.5
Hungary	–63.3	–67.2	–63.4
Israel	–73	–70.3	–51.5
Italy	–42.2	–43.5	
Norway	–40.8	–36.4	–30.2
⋮	⋮	⋮	⋮
United States	–53.6	–58.5	–36.1

Stacked Country, Year	*Spending Preference*
Australia, 1986	–45.3
Australia, 1996	–37.3
Australia, 2006	–12.8
Germany, 1986	–51.8
Germany, 1996	–67.5
Germany, 2006	–54.9
Great Britain, 1986	–7.9
Great Britain, 1996	–12.8
Great Britain, 2006	–4.5
Hungary, 1986	–63.3
Hungary, 2006	–67.2
Hungary, 2006	–63.4
⋮	⋮
United States, 2006	–36.1

purposes of analysis. For example, we could take individual-level total government spending preference data and calculate the net preference for total government spending for the individuals surveyed each month. This gives us a monthly time series of net preference for total government spending in Britain.

Time series, panel, and pooled cross-sectional data are all forms of longitudinal data. Before looking at the analysis of time series data, we need to specify some notation and terminology.

1.2 Time Series Notation and Terminology

One of the major stumbling blocks for students trying to understand time series analysis is the notation used. This is understandable, as reading any quantitative methods literature without knowing the notation is a bit like trying to read a text in a foreign language. Unfortunately, there is no single agreed-on notation, so the notation used in this text may differ from what you read elsewhere, but the notation is consistent throughout this text and the supplementary material. We begin with some basic notation:

X, Y, Z, W	Variables
x, y, z, w	Some single value (element) of the variable (i.e., the value of the variable at some unspecified single time point)
i, j, k, l, s, t	Indices (t is usually reserved to index time)
x_t, y_t, z_t, w_t	A specific value (element) of the variable (i.e., the value of the variable at some specified single time point)
T	Total number of observed time points: $t = 1, 2, 3, \ldots, T$

Some notation is specific to time series data. If Y is a time series variable, we often give it the subscript t: Y_t. The subscript t does not indicate that we are referring to a specific value of Y_t. It only indicates that Y is a time series variable. We will use y_t to denote the specific value of Y_t at some time point t. For example, y_1 or $y_{t=1}$ is the specific value of Y_t at the first time point; this is our first observation in the time series.

The Data-Generating Process Versus the Data Model

In the following chapters, we will often describe what we assume to be the data-generating process for the time series data that we are analyzing. This language may be new to many and needs some explanation.

When learning about simple linear regression, the distinction is often made between the *population model* or *true model* and the model estimated from the sample. In this language, the population model describes the process that generates the data from which we sample. For example, for variables X and Y, we may assert the population regression model as follows:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

$$\varepsilon \sim \text{NID}\left(0, \sigma_\varepsilon^2\right), X \sim \text{NID}\left(\mu_X, \sigma_X^2\right), E\left(\varepsilon \mid X\right) = 0. \quad (1.2.1)$$

Variable Y is a function of X and ε, which are themselves normally distributed random variables that are unrelated to each other (independent), denoted as NID. In this assertion, we assume that this is the process that generates the set of data that is our sample. Our sample represents N draws from this stochastic (containing a random component) process—specifically N random draws from X and ε, which then determine Y.[1] From the sample, we can specify the sample regression function as follows:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

$$Y = \{y_1, y_2, \ldots, y_N\}, X = \{x_1, x_2, \ldots, x_N\}. \quad (1.2.2)$$

This model, which we will estimate using our sample data (e.g., using ordinary least squares [OLS] estimation), is called the *data model* (sometimes called the *empirical model*). We indicate the model from a particular estimation using the "hat" notation:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\varepsilon}. \quad (1.2.3)$$

The data-generating process often is described as a stochastic function that could produce an infinite number of possible outcomes. Our data are N possible draws from this function. In many cross-sectional contexts, it is easy enough to think of the sample data as a random draw of N cases from the *very large* population of cases available for observation. The randomly selected cases provides us with our sample values of Y, X, and ε. In thinking about our data this way, it isn't really necessary

[1] Note that X may also contain nonstochastic elements.

to use the language of a data-generating process—although it may still be useful.

In the context of time series data and often in the context of cross-sectional data, our sample data are not a random draw from a population of cases available for observation.[2] Consequently, the language of a data-generating process becomes not just useful but necessary. In the time series context, the data-generating process is again described as a stochastic function that could produce an infinite number of possible outcomes. For example,

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \text{ for } t = 1, 2, \ldots, T$$

$$\varepsilon_t \sim \text{NID}\left(0, \sigma_\varepsilon^2\right),\ x_t \sim \text{NID}\left(\mu_X, \sigma_X^2\right), E\left(\varepsilon_t \mid x_t\right) = 0. \tag{1.2.4}$$

However, our data are a single draw from this data-generating process, in that we only ever draw one value of x_t and one value of ε_t for each time point $t = 1, 2, \ldots, T$. These then determine the values of y_t observed. The single draw of x_t and ε_t for each time point $t = 1, 2, \ldots, T$ is commonly called a single *realization* of the data-generating process. In time series analysis, we are not in a position to go back and resample different values of x_t and ε_t for a particular time point. Talking about the population of cases we could have observed at a single time point is meaningless, and so we instead talk about the stochastic process that generates the single value at that time point.

When describing fundamental time series concepts, we will often define the data-generating process that corresponds with each of the concepts. When describing the application of time series analysis, we will discuss the consequences of different data-generating processes for the model we estimate from our data (the data model). It will become evident that it is not always necessary for the estimated data model to contain all of the elements of the assumed function defining the data-generating process. We may assume that the data-generating process is as follows:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + \varepsilon_t, \text{ for } t = 1, 2, \ldots, T$$

$$\varepsilon_t \sim \text{NID}\left(0, \sigma_\varepsilon^2\right),\ x_t \sim \text{NID}\left(\mu_X, \sigma_X^2\right),\ z_t \sim \text{NID}\left(\mu_z, \sigma_z^2\right),\ E\left(\varepsilon_t \mid x_t, z_t\right) = 0. \tag{1.2.5}$$

[2] For example, when the cases in the cross-sectional data are European countries, we never really have a random draw of these countries.

At the same time, we might explore the consequences of this data-generating process for an OLS estimation of a data model of the following form:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \text{ for } t = 1, 2, \ldots, T \quad (1.2.6)$$

It is usually the case that we believe the data-generating process is more complicated than the data model estimated. It is not uncommon to be aware that y_t is, in part, determined by z_t without having any direct measure of z_t. The data-generating process can be interpreted as the unobserved reality that we are trying to reveal with our analysis. However, much analysis is focused on trying to reveal only a part of the data-generating process, while guarding against the possibility that the larger reality might lead us to reach false conclusions.

In time series analysis, as in other forms of statistical analysis, there is often an iterative process between the data-generating process assumed and the data model estimated (Hendry, 2003, chap. 1). Typically, we begin by stating the assumed data-generating process. This informs the data model that we then estimate. The results from the estimated data model may provide confirmation that our data-generating assumptions are correct or may contradict those assumptions. Accordingly, we may adjust our assumptions regarding the data-generating process and estimate a new data model. This process continues until we feel that there has been a convergence between our assumptions regarding the data-generating process and the data model estimated from our sample data.

The Lag of a Time Series Variable

Continuing on with time series notation, we use the notation y_{t-1} to indicate the specific value of Y at the time point just previous to t. We call this the *lag* of y_t. Say our observations are weekly; then, y_{t-1} at $t = 5$ (Week 5) is equal to $y_{5-1} = y_4$. This is the value of Y_t in Week 4. For clarity, this is sometimes called the *first* lag. Similarly, y_{t-2} is called the *second* lag. This can, of course, be applied to any variable (e.g., X_t or Z_t).

To "lag" a variable is to create a new variable where the value of the variable at a given time point t is replaced by the value of the variable from the previous time point, $t - 1$. Consider the following data (Table 1.4) containing the variable "social program spending preference" (measured on the same scale as the "total government spending net preference" variable in the previous example) in the third column. The lagged value of social program spending preference would look as it does in the fourth column. If we are modelling social program spending preference

Table 1.4 Social Program Spending Preference—Lags, Leads, and First Differences

Year	*Unemployment*	*Social Program Spending Preference*	*Social Program Spending Preference —First Lag*	*Social Program Spending Preference —First Difference*	*Social Program Spending Preference —First Lead*
1988	7.71	27.45	—	—	32.32
1989	7.59	32.32	27.45	4.87	25.18
1990	8.84	25.18	32.32	–7.15	22.39
1991	10.46	22.39	25.18	–2.79	20.26
1992	11.35	20.26	22.39	–2.13	15.53
1993	11.38	15.53	20.26	–4.74	20.63
1994	10.08	20.63	15.53	5.10	23.80
1995	9.51	23.80	20.63	3.17	33.41
1996	9.71	33.41	23.80	9.62	40.13
1997	8.95	40.13	33.41	6.72	45.08
1998	8.15	45.08	40.13	4.95	46.69
1999	7.30	46.69	45.08	1.61	44.89
2000	6.90	44.89	46.69	–1.80	—

Note that in practice you lose a time point when lagging a variable. In discussions of the theory of time series analysis, this is sometimes overlooked, but it will be important to keep in mind when we discuss the practical application of time series analysis.

and we include a lag of this variable as an explanatory variable, this variable is often called a "lagged dependent variable," as it is just that—the lagged value of the dependent variable. More will be discussed on this in Chapters 2 through 4.

The Lead of a Time Series Variable

To "lead" a variable is to create a new variable where the value of that variable at a given time point *t* is replaced by the value of the variable from

the *next* time point, $t + 1$. The lead value of "social program spending preference" would appear as it does in the last column of Table 1.4. The notation for the lead of y_t is y_{t+1}. It is rare that such a variable would be included as an explanatory variable in a model. It is unlikely that we would expect the future value of the dependent variable to be an explanatory variable, but in Chapter 6 we will see an example of its use.

The Difference of a Time Series Variable

To first difference a time series variable is to create a new variable where the value of that variable at a given time point t is equal to the value of the original variable minus the value of the first lag of the variable. The second-last column of Table 1.4 contains such a first difference of social program spending preference. The 1989 value of 4.87 is the result of subtracting the lag of social program spending preference (27.45) from the original 1989 value of social program spending preference (32.32). The notation for the first difference of y_t is Δy_t —the Greek letter *delta* is often used to indicate change or difference. The first difference of a variable can be interpreted as the change in this variable since the previous time point. The change in social spending preference from 1988 to 1989 is 4.87 (an increase).

Getting a grasp on notation and terminology is one of the greatest hurdles to understanding time series analysis. This text will build on the notation and terminology outlined in this chapter as needed, but what has been presented so far provides the framework necessary for the discussion of time series fundamentals in Chapter 2. Before we move on to those fundamentals, let us discuss some of the key opportunities and challenges presented by time series analysis.

1.3 Opportunities and Challenges With Time Series Data

To explore the problems created by time series data for the methods of analysis generally used with cross-sectional data, let us consider the following example. For this example, we will analyze data collected for the purpose of testing the thermostatic model of public responsiveness and policy representation as developed by Soroka and Wlezien (2010).

The thermostatic model is actually two models that describe how (1) public demands for increases or decreases in policy spending respond to current levels of government spending and (2) government changes in policy spending respond to public demands for increases/decreases. The following examines the first of these, called a public responsiveness

model. The unit of time for this model is the fiscal year—there is one observation per fiscal year. The public responsiveness model is as follows:

$$R_t = \beta_0 + \beta_1 P_t + \beta_2 W_t + \varepsilon_t, \tag{1.3.1}$$

where R_t is the public's relative preference for policy spending in a given year—that is, the difference between the public's preferred level of policy spending and the level that it actually gets.[3] P_t is the actual level of policy spending in a year. W_t represents other, exogenous effects on the public's relative preferences—this could include more than one variable. For our current purposes, let us just regress the public's relative preference for policy spending on the actual level of policy spending (in millions of Canadian dollars). We do this using yearly data on the Canadian public's relative preference and government policy spending for social welfare payments from 1988 to 2003. Table 1.5 contains the results from an OLS estimation.

In addition to estimating the intercept and slope coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$, and calculating the corresponding t statistics, we usually estimate a goodness-of-fit statistic (e.g., R^2).

Recall that we typically use a t test to test the statistical significance of the regression coefficients. Looking at Table 1.5, if we were using a 0.05 significance level, we would conclude that we could not reject the null hypothesis that the slope coefficient for program spending is 0. Therefore, we could not reject the null hypothesis of no effect for program spending on relative spending preference. If we were using a 0.10 significance level, we would conclude that there is a statistically significant and positive relationship between government social welfare spending and the public's relative preference. For each additional billion dollars spent on

Table 1.5 Canadian Public's Response to Government Spending

Preference	*Coefficient*	*Standard Error*	*t Statistic*	*P Value*
Program spending	0.32	0.16	1.98	0.067
Constant	–25.37	30.03	–0.84	0.412

NOTE: $R^2 = 0.22$, $T = 16$; T = number of time points.

[3] This is operationalized using the measure described in Table 1.2.

social welfare programs, the public's relative preference increases by 0.32 on the (−100, 100) scale. This suggests that greater spending leads to an increase in the demand for spending! Neither of these is what Soroka and Wlezien (2010) predicted, but this is also not the model they used—and for good reason.

The preceding analysis included a number of assumptions. One of the most important of these assumptions is that the observations are independent. Usually, the cases in a cross-sectional data set are assumed to have been selected randomly, and therefore, the value of any case for a given variable will be independent from the value of any other case for the same variable. In time series data, we have measures of the same variable for the same single case at different time points. Therefore, time series data, such as those used here, are usually not independent, especially if the sampling time interval is small. Observations close together are often more alike than those far apart. For example, public opinion today is more closely related to public opinion yesterday than it is to public opinion last year. In our current example, this is public opinion regarding government spending levels. Addressing the potential for time series data to violate the assumption of independence motivates many of the analytical approaches discussed in this text. For now, let us consider some of the consequences of this violation.

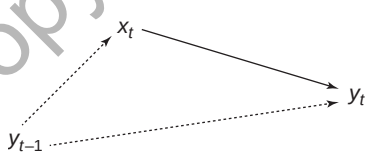
If not accounted for in our analysis, one of the problems the violation of independence can lead to is a problem called serial-correlated errors. This is the problem of correlation across the estimated errors in our data model. We discuss this further in Chapters 2, 3, and 5. For now, consider the possibility that the effects captured by the error term at one time point may be correlated with the effects captured by the error term at the next time point. Relating back to our example, this could happen if an event captured by the error term one month increases the public's wish for spending and, at least, part of this effect remains and is captured by the error a month later.

A second potential problem is as follows. A source of the lack of independence may be that public opinion today is partly explained by public opinion yesterday. If we are modelling public opinion on a daily basis and public opinion yesterday is a predictor of public opinion today, this predictor should be included in the model. This predictor is the lag of the dependent variable discussed earlier. The substantive interpretation of including such a variable in a model will be discussed in Chapters 2 and 3. For now, it is sufficient to note that not including it under these circumstances can lead to a violation of another important assumption of OLS regression. This is the zero conditional mean assumption. Let us state this assumption in terms of our current example:

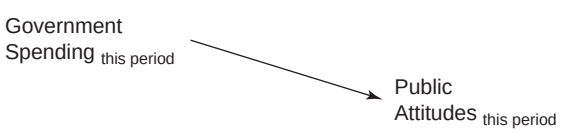
$$E(\varepsilon_t \mid P_t) = 0. \tag{1.3.2}$$

This implies that the (conditional) mean of the error term is independent from the explanatory variable(s). What if the past year's relative policy spending preference is a predictor of the current year's relative policy spending preference (R_t), as suggested above? If it is also a predictor of our explanatory variable—current policy spending (P_t)—Assumption (1.3.2) will be violated. This may be the case if government spending levels are a function of the public's relative preference in the previous fiscal period. This is very similar to the omitted-variable bias problem familiar to those who have studied cross-sectional data analysis. It is called an endogeneity problem and is discussed further in Chapter 2.

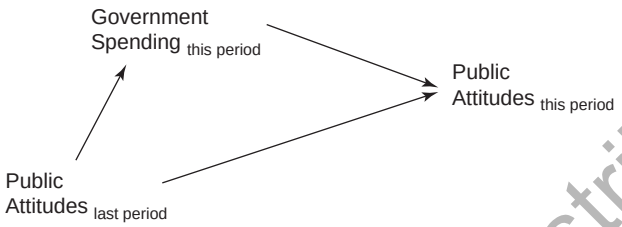
Those who have studied cross-sectional data analysis will be familiar with the problem of omitted-variable bias. This is a particularly pernicious problem as there is no direct way of testing for omitted-variable bias except for including the variable(s) that is(are) suspected to be omitted. The difficulty that often arises is that a variable that is commonly suspected to be omitted is the past value of the dependent variable. In the figure below, we may be interested in the effect of x_t on y_t, but we are concerned about the effect of y_{t-1} on both x_t and y_t in the data-generating process. In other words, we are concerned that the omission of y_{t-1} from the data model might produce an omitted-variable bias.



This concern is actually relatively common. Consider this problem in the context of the current example. We regress public attitudes toward government spending levels in a given fiscal period on actual government spending in the same fiscal period:



An omitted-variable bias will occur if public attitudes toward spending in this fiscal period are correlated with public attitudes in the previous fiscal period and if those attitudes in the previous fiscal period affect government spending this fiscal period.

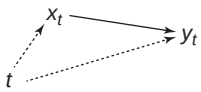


This is one of the more common violations of the zero conditional mean assumption, known as the exogeneity assumption. This assumption is discussed further in Chapter 2. As we will see, time series analysis allows us to test and correct for the problem by including the past value of the dependent variable directly in the data model. This is one way in which time series data provide us with an opportunity we do not have with cross-sectional data. We will explore other opportunities throughout this text.

The analysis of time series data also introduces additional problems. For example, another potential violation of the zero conditional mean assumption is that both x_t and y_t are trending. A variable trends if, in addition to other dynamics and random variation, it increases or decreases by a constant magnitude each time period. This could occur for a number of reasons, such as if social welfare programs have steadily become more expensive to provide over time and the public's expectations regarding the provision of those programs have also increased steadily. Let us visually examine the variables: the public's relative preference for social welfare spending and actual social welfare spending (Figure 1.1).

Clearly, both variables are trending upward. If two series are trending together, we will probably estimate a strong correlation between the two, but we can't assume that the relation is causal. An alternate possibility is that it is a spurious result produced by the fact that both variables are a function of time. The difficulties with and approaches to trending are discussed further in Chapters 2, 3, and 6.

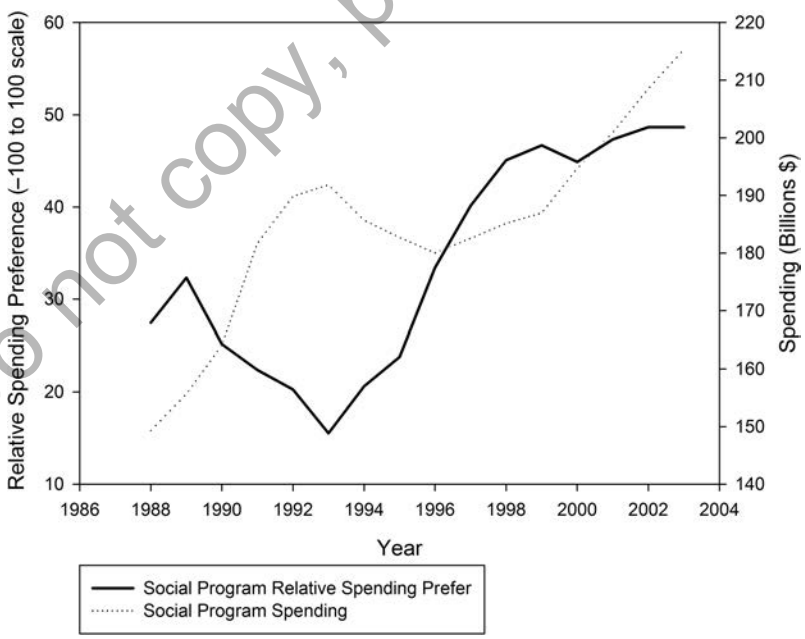
Generally, if we are interested in the effect of x_t on y_t, we need to be concerned if both x_t and y_t appear to have data-generating processes that are a function of time. If both x_t and y_t trend and this is not accounted for in the data model, our estimation of the effect of x_t on y_t will be subject to a spurious correlation, akin to omitted-variable bias.



As another concrete example, it is quite common for a new government's popularity to begin trending downward after an initial honeymoon period. This could be driven by any number of things: The new government's popularity was artificially high due to the positive coverage from the election win; once the government has to start making decisions it inevitably upsets some supporters; and so on. As a consequence, any other variable that trends downward or upward during the same period will correlate significantly with government popularity even if it is not related to it in any way. This will hold for other functions of time as well, as we will discuss in Chapter 3.

Trending also violates an assumption that is unique to longitudinal (e.g., time series and panel) data. This is the assumption of stationarity. The nature of the data-generating process of time series data discussed earlier in

Figure 1.1 Canadian Public's Relative Preference and Actual Government Spending

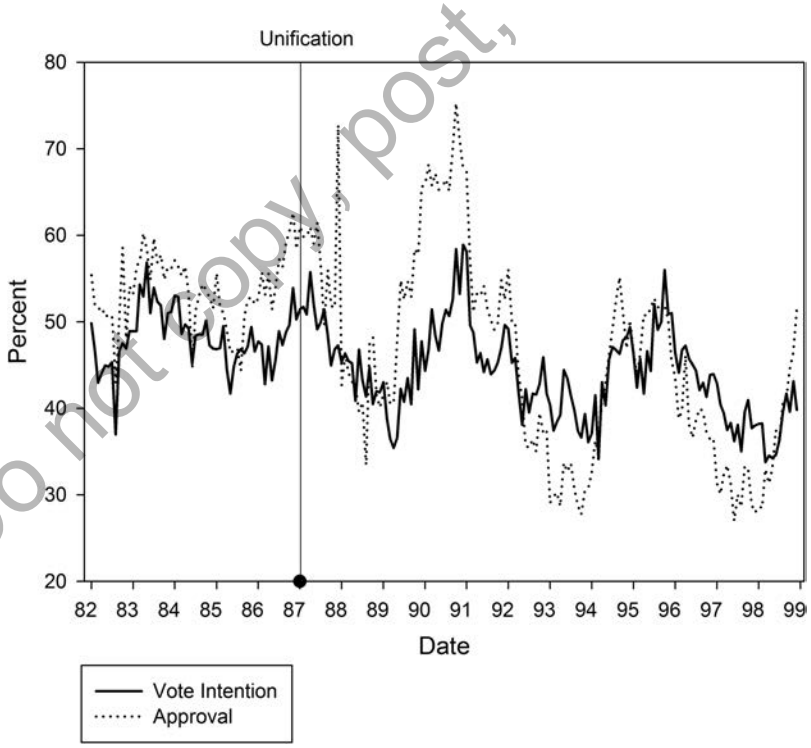


this chapter will require us to make this complex assumption. This is discussed further in Chapter 2.

A final challenge presented by time series data that we have not yet touched on can be illustrated by another example. We may be interested in how approval ratings for the German government translate into vote intention for the government over the period from 1982 to 1998 (Figure 1.2). Over this period, our data are for the government of West Germany prior to January 1987 and for the government of the unified Germany subsequently. If we plot our approval and vote intention time series, we will note something important.

After the reunification of Germany, vote intention becomes somewhat more volatile and approval even more so. This is likely due to the weakening of partisan identification in the latter period (Pickup, 2010). What we are seeing is a structural break in the variances of both time series. This

Figure 1.2 German Government's Approval and Vote Intention



structural break will present a problem for any model that assumes that the variances in the time series are constant across the period of analysis. Structural breaks can also occur in the means of time series and the covariances between time series. In fact, it is quite likely that the covariance between approval and vote intention also has a structural break in January 1987. We will need to account for such breaks in our time series models. As we will see in Chapter 5, such structural breaks will sometimes be a nuisance to be dealt with but at other times they will be of interest in and of themselves.

Summary

In this chapter, you have been introduced to some of the basic notation and terminology of time series analysis. You have also been given a taste of the opportunities and challenges presented by the analysis of time series data. In the following chapters, we will explore these and other opportunities and challenges further. We will learn how to address the challenges and take advantage of the opportunities. In the next chapter, we continue our introduction to time series analysis by surveying the fundamental concepts of time series data and analysis.